テキストマイニングによる「未来社会」解読の試行

■ 論　文 ■

テキストマイニングによる「未来社会」解読の試行

原　　以　起[*1]・奥　野　圭太朗[*2]
奥　野　卓　司[*3]

■ 要　　旨 ■　「未来社会」の予測を、従来の社会科学的方法、社会調査による方法以外に、大量のテキスト・データの解析によって行うことは、果たして可能であろうか。筆者らは、テキストマイニングのソフトウエアを使う方法が、どの程度、社会科学として有用性がありうるのかを検証するために、実際にそれを試験的に試み、この方法に一定の有用性を確認することができたので、本稿で報告する。

本稿でテキストマイニングの対象としたのは、小松左京の SF 作品『虚無回廊』である。これは、この作品が、ある程度科学的根拠にもとづいて創作されており、作品中に多用されている「AI（人工知能）」および「マン・マシン・インタラクション」が将来的に社会に大きな影響力をもつであろうと推定されたからである。このため、この 2 技術に関連したキーワードを核に、この作品をテキストマイニングし、この作品内での社会におけるその技術の位置づけ、人間との関係性を抽出し、その紐帯図を KJ 法とブレーンストーミングにより検討することで、「未来社会」のありようを解読した。

その結果、今回は社会科学的な「未来社会」像の提示ということでは不十分と言わざるをえなかったが、この研究方法自体は間違ってはいないと判定できた。本稿では、その「方法論」を提示する。

■ キーワード ■　テキストマイニング、未来社会、社会調査法

*1　ヤマハ発動機（株）ボート事業部先行開発部 470 プロジェクトグループリーダー
*2　熊本学園大学商学部特任講師
*3　関西学院大学社会学部教授／先端社会研究所所長／（公財）山階鳥類研究所所長

1. 研究の目的

私たちはどこから来たのか、どこに行くのか、ということは、人類の永遠の問いである。したがって、未来をどのようにつかむかを、人間は呪術、宗教、哲学から暦学、確率論、経済予測などとしても、それに挑んではきた。

だが、一方で、社会科学では「未来」は扱ってはならないことになっている。社会科学が科学である以上、過去の事実を基礎にした理論から論理的に展開しており、それをいかに延長しても、現在より先につながることはない。また、社会調査はその調査時点の「現在」を語っているのであり、「未来」を語らない。その意味で、社会科学が「未来」を扱わなかったのは、学問的には適切だったと言えよう。

しかし、一方、学問が現実の社会に貢献しようとすれば、その実践は「現在」から「未来」に向けてなされるべきであるということは、多くの研究者が考えているところだ。その「未来社会」の姿を明らかにし、そのうえであるべき実践をおこないたいという、彼らの知的欲求もまた抑制できるものではない。

そこで、「確率」、「傾向」、「周期」など、いくつかの「科学的方法」に近似する方法によって、未来の予測が行われてきた。しかし、社会は連続的に展開していくのではないことは、人類の歴史から明らかである。確率論からの推定が一度しかない「未来」の予測に確実な方法ではありえない。

このように科学的に「未来社会」を予測することはほぼ不可能である。しかし、その時点での、相対的に納得できる方法でできうるかぎりその近似値を提示することは、その「未来社会」を人類の多様な幸福に結びつけていこうとする社会的実践、とりわけ社会科学の寄与に意味があると言えるのではなかろうか。

一方、「未来社会」を従来の社会科学的方法、社会調査による方法以外で、仮想的に描いてきた文芸として SF がある。もとより SF は作家によって書かれた「物語」であり、実証性はない。しかし、SF のなかには、科学的根拠にもとづいて創作された作品群があり、それらはあくまで文芸であるため、論理では訴求できない飛躍によって、かえって深層が描かれていることがある。そして、この深層は、それらの作品を社会科学的な手順にしたがって分析することで、発掘することも可能である。

そこで、これを発掘する一つの社会科学的方法として、そうした SF の中で使用されている「技術的術語」のうちから、未来を拓く可能性のある技術とされているものに着目し、その術語を核に、複数の作品群の全文章をテキストマイニングして、その作品内での社会における、その技術の位置づけ、人間との関係性を抽出することで、「未来社会」のありようが見えてくるのではないかと筆者らは考えた。

同様ことは、過去にも考えられなかったわけではないが、大量のテキスト・データを使うため、数年前までのパーソナルコンピュータの処理能力では極めて困難だった。しかし、今日では、コンピュータの処理能力が飛躍的に向上するともに、テキストマイニングのための優秀なソフトウエアが開発されたため、比較的容易に可能になった。

この方法が、どの程度、社会科学として「未来社会」を予測することに有用性がありうるのかを検証するために、実際にそれを試験的に試みることによって、この方法の有用性を確認する一定の知見をえることができたので本稿で報告する。

2. データ分析方法

上記の目的で、本稿では、テキストマイニングの対象とする技術術語として、今日、注目されている科学技術のうちで、将来も社会的に重要となるものという観点から、「AI（人工知能）」と「マン・マシン・インタラクション」に関連する術語をキーワード（後述）として採用することにした。

また、マイニングの対象とする SF 作品の作家候補として、小松左京、星新一、筒井康隆、アーサー・C・クラーク、マイケル・クライトンらの作品を検討した。

これらから内容の哲学的概念の質、技術と人間との関係に関する考察の質とともに、その作品のデジタルでのテキスト・データの入手・使用可能性を検討した結果、上記の 5 名の作家中では、小松左京（1931～2011）が長編 17、中短編 269、ショートショート 199 の作品を創作し、それらの作品に上記の要素が強く、かつデジタル・テキスト・データ化されていることが判明したため、今回は小松左京の作品によって、この研究を試みることとした。

先述したように、今回は、あくまでこの方法が、「未来社会」の社会科学的な予測に一定の有効性をもちうるかどうかを、実験的に知ることが目的であるため、小松左京のすべての作品を一挙に対象とはせず、その事例として、小松の作品の中で、『復活の日』、『果てしなき流れの果に』、『継ぐのは誰か』、『虚無回廊』の 4 作品のうちから、どの作品を今回のマイニングの対象とするか、慎重に検討した。

この経過を、小松左京作品の著作権をもつ小松家の方に相談し、小松家から本稿の意図にもっとも相応しい作品として『虚無回廊』を推薦され，その使用を無条件に承諾されたため、同作品を対象として、テキストマイニングを行うこととした。そして、この方法に一定の有効性が見いだせれば、続けて他の 3 作品を、今回と同様の方法、もしくはそれに修正を加えた方法で行うことに、両者が同意した。

『虚無回廊』は、小松左京が 2011 年に逝去しため、未完となった作品で、人間と AI のハイブリッドである AE（人工実存）を物語のキーとして、究極のネットワーク社会が描かれている。

この作品を対象として、すでに決定していた「AI（人工知能）」と「マン・マシン・インタラクション」に関連するキーワード間の共起関係を紐帯図で可視化する。

これにより、

① キーワード間の関係の強さの可視化――何が近いとみなされ／何が遠いとみなされているか

② キーワード間を繋いでいる共通項の可視化――異なるキーワード同士の共通性

③ キーワードごとの共起単語の可視化――どのようにキーワードは語られているか

を明らかにすることとした。

この過程は表 1 に示すように、文章を単語単位に分解し、そのなかで共起しやすい単語を演算

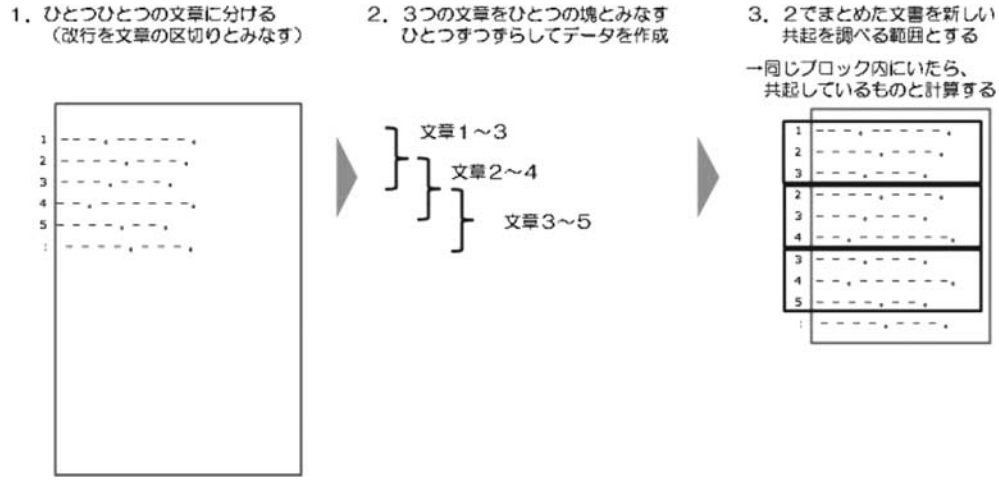
し、キーワード間、及びキーワードとその共起単語の「関連強度」（後述にて解説）を紐帯図で可視化する（図 1）というプロセスで構成される。

表 1　テキストマイニングの過程

①分析用データを作成	**共起関係を探るうえで、ある単語の周辺をみる範囲を少し広めに設定** ある単語と関係の深い単語は、ある程度位置が離れても出現すると仮定 そのため、改行を文章の区切りと見なした上で、３つの文章をひとつの塊と見なし、その中に含まれている場合、共起しているとみなした
②単語単位に分解	**共起関係をみるため、文章を単語単位に分解** 単語は、名詞、形容詞、動詞に絞った（記号や接続詞などは除外）
③注目ワードを設定	**小説で関係性を知りたい単語を選定** 機械で設定するのではなく、人間が目検で探索し、設定
④注目ワードと共起しやすい単語を演算	**注目ワードと共起しやすい単語を情報量で抽出** 注目ワードの周辺にある単語の共起しやすさを情報量で計算 注目ワードと関係の強い単語を抽出した
⑤ネットワーク図で可視化	**③で抽出した単語をネットワーク図で可視化** 注目ワードと、関係の深い④で抽出した単語を線で結び、ネットワーク図を作成 情報量上位●件までや、情報量●以上などといった足切りのパターンを変えて可視化

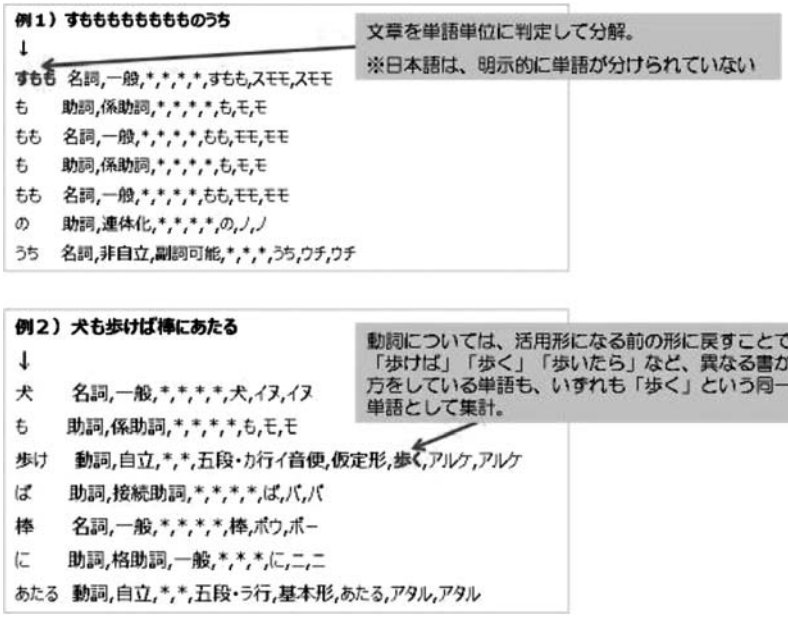
①分析用データの下処理（1）

分析用データは、キーワードを共起する文言が必ずしも 1 センテンス内に存在するとは限らないため、共起する単語が少し離れていても共起とみなすよう、ある程度幅をもった形でブロックし、1 センテンスづつスライドしていくようにインプットした。



②分析用データの下処理（2）

文章を単語単位に分解し、集計可能な状態にした。



③「キーワード」の設定

先端技術として注目されているもの、人間の振る舞いに関するもの、環境に関するものから、以下のキーワードを抽出・設定した。

技術ワード

- 人工
- AI（人工知能）
- AE（人工実在）
- ロボット
- VP（ヴァーチャル・パーソナリティ）
- センサー
- ネットワーク
- バイオ
- ニューロン

人間ワード

- 性
- コミュニケーター（orコミュニケーション）

環境ワード

- 中枢
- 生物
- 森
- 種子
- 都市

④「キーワード」と共起しやすい単語を演算……関連強度＝「情報量」の算出

情報量は、その定義から、全体に比べてその分布の偏りがどれくらいの確率で起こるかを計算しその対数をとって表した数値のことである。「出現確率が低い」ほど、情報量は大きくなる。

	単語Aが含まれる	単語Aが含まれない	合計
注目ワードXが含まれる	両方含む文章数：67	注目ワードXのみの文章数：363	430
全体	200	2085	2285

情報量

$$I = -\log P$$ ←確率Pをマイナスの対数で表わす※

$$= -\log\left(\frac{{}_{200}C_{67} \times {}_{2085}C_{363}}{{}_{2285}C_{430}}\right)$$ ←単語A、200文章が含まれる2285の文章データの中から、ランダムに430を抜き出した時、その中に技術用語Xが67含まれる確率を計算。

$$= -\left(\log\frac{200!}{67!\cdot 133!} + \log\frac{2085!}{363!\cdot 1722!} - \log\frac{2285!}{430!\cdot 1855!}\right)$$ ←分解した計算式

$$= 16.23$$

本稿では、確率のまま扱わず、マイナスの対数で表わしたが、それは桁が膨大になるため、対数の方が数字の見た目が扱いやすいからである。これによって、0.00000000000*****%という確率も 2 ケタ程度の正の値で表現できるため、見やすくなる。また対数の底は、同じ理由で自然対数の 10 を使用した。

⑤情報量数値の目安

本稿では、情報量数値の目安として、

3 以上あれば、やや確からしい数字　⇒　特徴が見られる

5 以上なら、起きにくい数字　⇒　特徴がある

20 以上なら、偶然ではまずありえない数字　⇒　かなり特徴的である

として、あつかった。

↓情報量

情報量 =-log(確率)	この分布の偏りが起こる確率は？	この分布の偏りが起こるとしたら、何回に1回の確率か
1	36.788%	3回に1回
2	13.534%	7
3	4.979%*5%検定だとここのあたり	20
4	1.832%	55
5	0.674%	148
6	0.248%	403
7	0.091%	1,097
8	0.034%	2,981
9	0.012%	8,103
10	0.005%	22,026
20	0.000000206115362243 86%	485,165,195
30	0.000000000009357622 97%	10,686,474,581,525
40	0.000000000000000042484%	235,385,266,837,020,000
50	0.000000000000000000002%	5,184,705,528,587,070,000,000

⑥紐帯図の作成

以上のようにして算出された「情報量」を、各キーワード間、及び各キーワードに共起される単語との「距離」に置き換え、全てのアイテムが 2 次平面上に並ぶように自動配置し、図 1 のように可視化した。

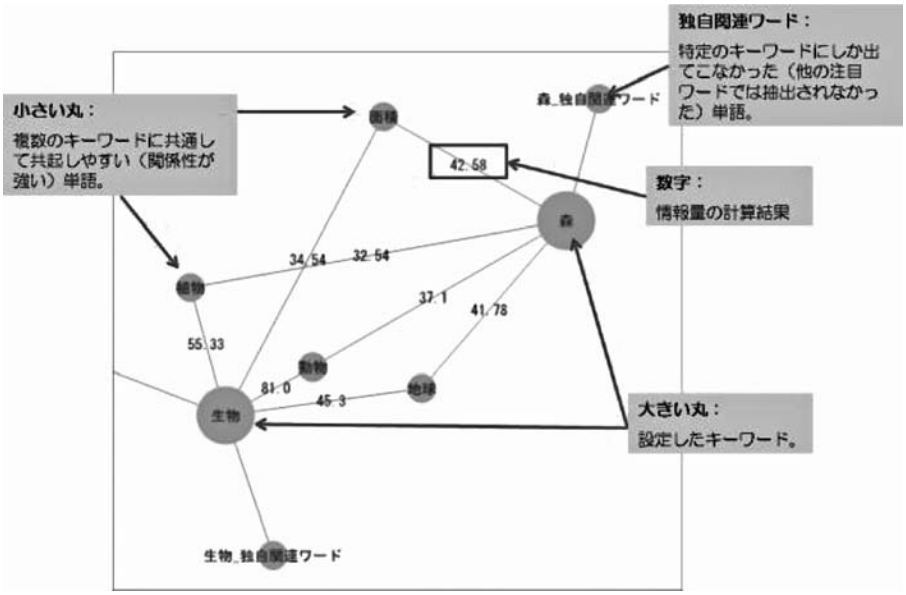
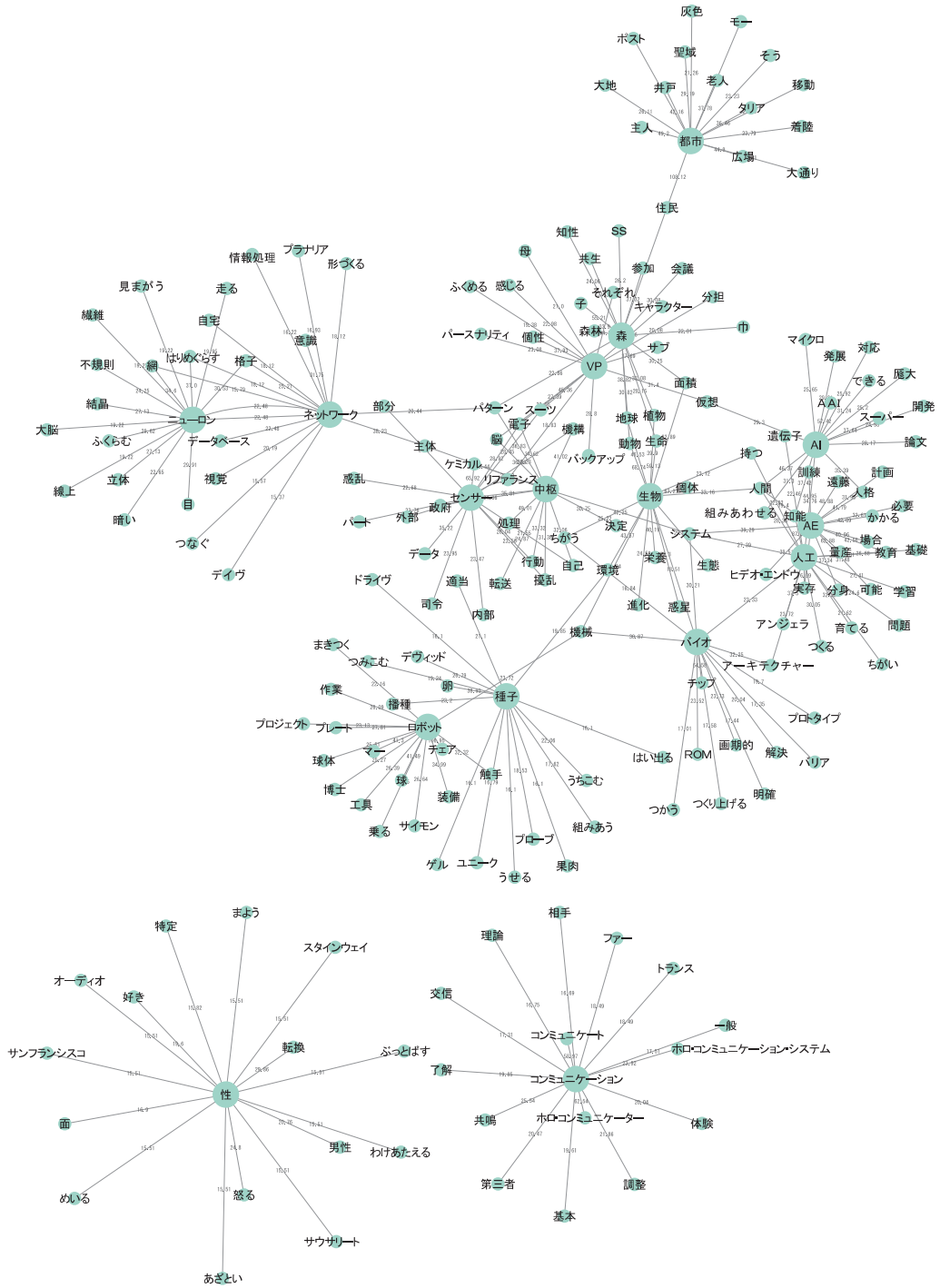


図 1　可視化作業のイメージ

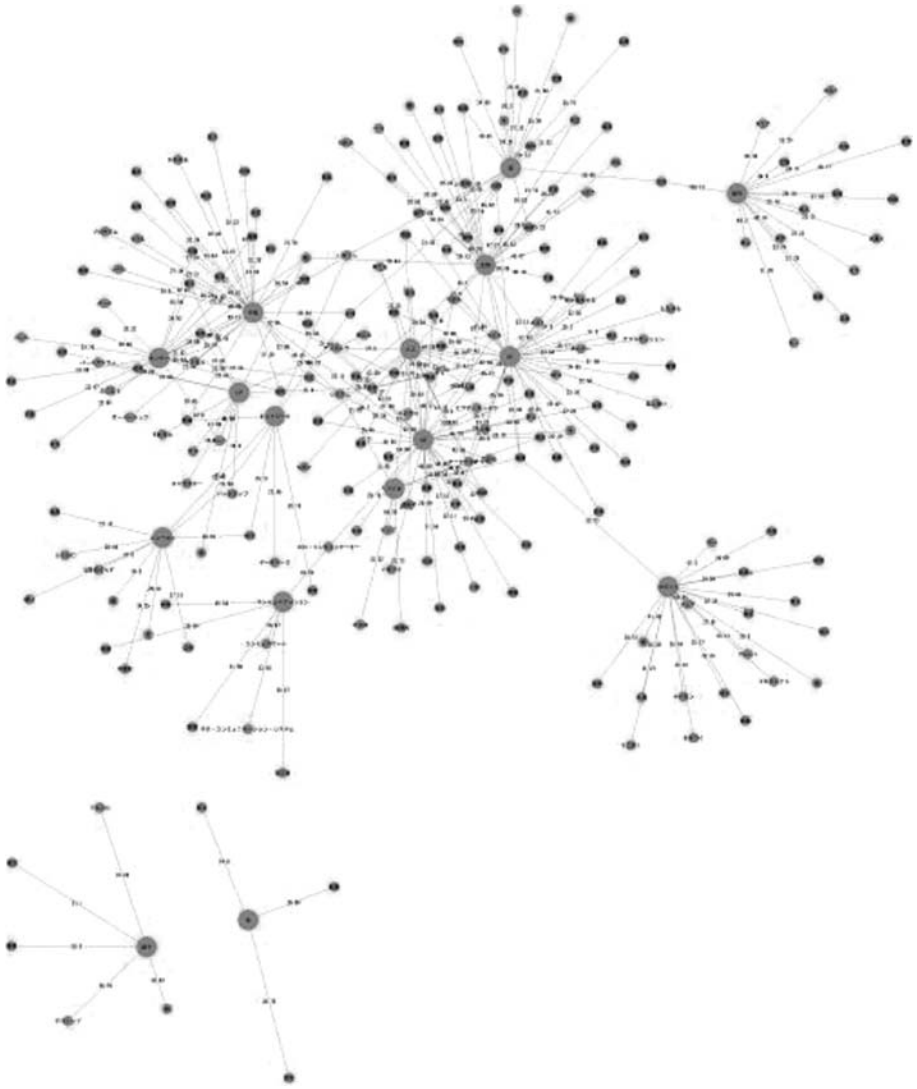
⑦「情報量」閾値毎のマッピング

⑤項で算出された「情報量」の、どのレベルまでをマッピング対象とするかによって、マップの様子が変わってくるため、マッピング時の情報量採用順位を 15 位まで、20 位まで、30 位までの 3 パターンで実行してみた。その結果、以下のような 3 種の紐帯図が得られた。

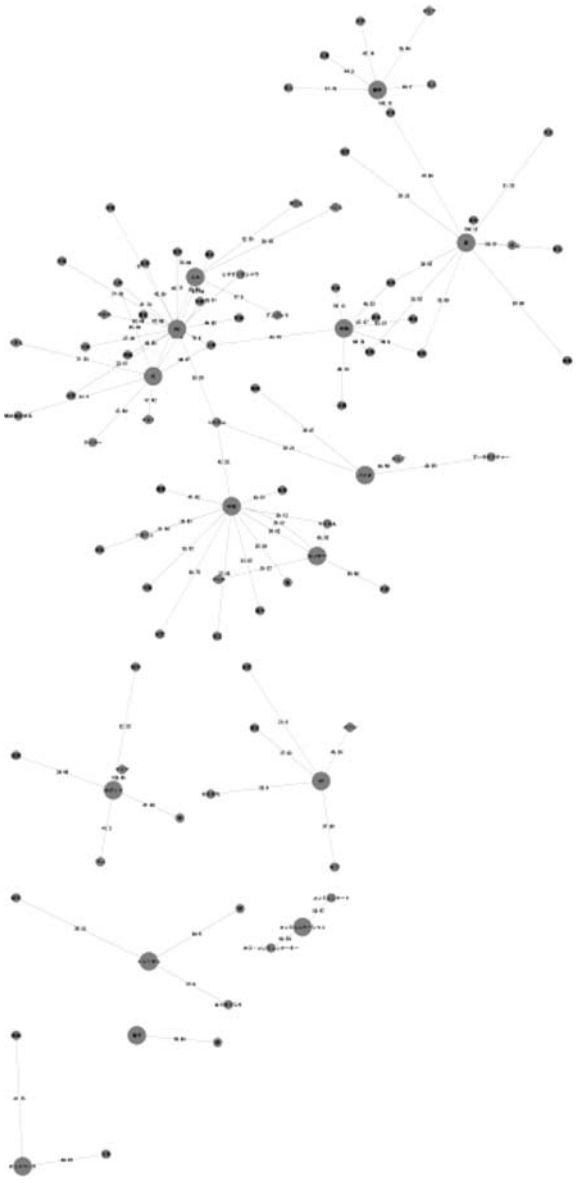
A：15 位まで


灰色
モー
ポスト
聖域
そう
大地
井戸
老人
移動
タリア
主人
着陸
都市
広場
大通り
住民
知性
SS
母
共生
参加
会議
ふくめる
感じる
それぞれ
子
分担
キャラクター
パースナリティ
個性
森林
森
巾
サブ
VP
面積
仮想
マイクロ
発展
対応
できる
AAI
開発
スーパー
論文
遺伝子
AI
持つ
訓練
計画
遠藤
人格
人間
個体
生物
組みあわせる
知能
AE
必要
かかる
場合
教育
基礎
人工
量産
ヒデオ•エンドウ
実存
分身
可能
学習
アンジェラ
育てる
問題
つくる
ちがい
バイオ
アーキテクチャー
チップ
ブロトタイプ
ROM
画期的
解決
バリア
はい出る
つかう
つくり上げる
明確
機械
進化
惑星
システム
決定
ちがう
行動
自己
環境
栄養
生態
ヒデオ•エンドウ
中枢
センサー
リファランス
ケミカル
感乱
政府
外部
処理
データ
転送
撹乱
適当
ドライヴ
司令
内部
脳
電子
機構
バックアップ
動物
生命
地球
植物
パターン
スーツ
部分
ネットワーク
主体
ニューロン
大脳
ふくらむ
データベース
立体
視覚
目
暗い
つなぐ
ディヴ
線上
結晶
繊維
見まがう
走る
自宅
不規則
編
はりめぐらす
格子
情報処理
プラナリア
形づくる
意識
まきつく
つみこむ
デヴィッド
卵
播種
種子
作業
プロジェクト
プレート
ロボット
球体
マー
チェア
触手
うちこむ
博士
球
工具
装備
乗る
サイモン
ブローブ
組みあう
ゲル
ユニーク
うせる
果肉
まよう
特定
スタインウェイ
オーティオ
好き
サンフランシスコ
転換
ぶっとばす
性
面
男性
わけあたえる
めいる
悠る
サウザリート
あざとい
相手
理論
ファー
トランス
交信
コンミュニケート
一般
ホロ•コンミュニケーション•システム
了解
コンミューニケーション
共鳴
ホロ•コンミュニケーター
体験
第三者
調整
基本

B：20 位まで



C：30 位まで



以上の 3 つの紐帯図を比べてみると、15 位までのものがアイテム間の距離が全体的に適度に広がりをもっているため、キーワード間の連関性や説明単語を視覚的に理解しやすいように判定できる。これに対して、20 位までの紐帯図は、キーワード間の距離が短く、全体的に過密な状態となってくる。さらに 30 位までとなると、キーワード間の連関性が希薄になりすぎてしまう。以上の比較から、20 位まで、30 位までは、本稿の作業には不向きであると判断し、最終的に 15 位までの紐帯図を採用することとした。

⑧紐帯図解読と発想（情報量 15 位までを採用した紐帯図による）

紐帯図の解読と発想にあたっては、キーワード同士の距離感や、それぞれを説明している特徴的な単語から、各キーワードがどのような機能をもちうるのか、それらがどのような関係性と価値を提示しているのかを、現段階で筆者らが知りえた最先端技術情報と照らし合わせながら、相応の蓋然性を意識しつつ、KJ 法による整理を行った。

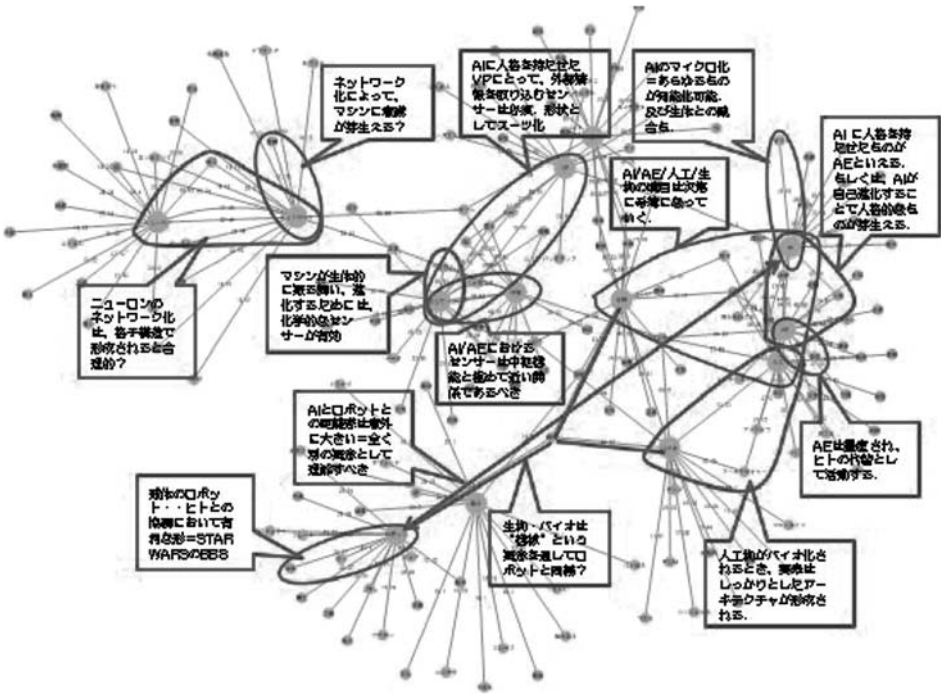


図 2　テキストマイニングによるキーワードの紐帯図

結果的に「テキストマイニングによるキーワードの紐帯図」は図 2 に集約できるであろう。この図 2 を解読していく時、

・キーワード間の関連強度と位置

・キーワードの説明ワード

・作品原文脈による解釈補完

という観点に重点的に着目して、以下のようにデータを解読した。

3. 解読結果

以上のマップから、共同研究会でのブレーンストーミングによって得られた要点は以下の通りであった。

・紐帯図上で「センサー」は「中枢」と近接しており、「中枢」は「システム」を介して「人工」

と関連付けられていることが多い。
・「パターン」は、「中枢」、「ネットワーク」、「森」の中心的な位置で関連付けられていることが多い。
・「ネットワーク」は「ニューロン」との関連付けが複数のアイテムを介して表現されている。
・AI/AE の乗った宇宙船を「都市」、AI/AE 内の VP を「住民」、行き先である天体の場を「森」と表現している。
・文明的な「都市」生活から、自然の豊かな「森」（もしくは「海」）に出かけ、リフレッシュし、新たな刺激を得て好奇心を満足させるといった状況をイメージさせる表現が多い。
・自然の仕組みをそっくりそのままモチーフにし、そこに人工的アイテムをはめ込み「人工」と「自然」との融合観を表現しようとしている。
・「ロボット」は、「機械」を介して、「バイオ」、「生物」と関連付けられていることが多い。
・「ロボット」は、「触手」、「チェア」、「球体」、「工具」、「作業」、「まきつく」などのキーワードと関連されている。
・用途や目的に応じて、生き物のように臨機応変に形状や動作を変更することが可能なロボットの存在が表現されている。
・「人工」と「バイオ（生体）」の共通アイテムとして「アーキテクチャ」がある。
・「バイオ」にはその複雑なシステムを構成し運用維持するために、システムフォーマットとしての「アーキテクチャ」が存在している。
・「バイオ」には「ROM」、「チップ」が繋がっている。
・「AE」、「VP」、「森」はいずれも「情動」と関連させられている。
・「VP」は、「あざとい」、「だます」を介して、「性」と繋がっている。
・「AI」「AE」には、ケミカル・「センサー」が備わっており、匂いや感触をセンシングし、それ自身で情報処理もするが、それを遠く離れた観測者と共有することができる。
・「AI」「AE」の中に概念的に存在する「VP」は「スーツ」を介して「センサー」と繋がっている。
・「AI」「AE」には実体はないが、外部との身体感覚的接点機能をもたせることで、「VP」の存在感にリアリティをもたせている。
・「性」に関連するキーワードは多様であるが、それらは本来人工物である「VP」（実質 AI）に生物的存在感をもたせる重要なアイテムである。
・「都市」、「生物」、「中枢」は、「恐怖」を中心にした位置関係にある。
・「AI」、「AE」、「バイオ」、「VP」、「センサー」、「性」に関連されている感情を表象するキーワードは比較的多様であり、まとまった位置にある。

以上の、紐帯図上のキーワードの相互関係を把握したうえで、それをもとに KJ 法により解読した結果は以下の通りである。

・高度に進歩した AI は、ディープ・ラーニングによる情報処理能力のみならず、人間の感情や感

覚をも実装するようになる。
・人間と同様のキャラクターを、機械で再現することが可能となり、身体が存在するかどうかとは別に、「もう一人の私」を仮想的に実在させることができる。
・AI、AE には、複数のキャラクター（人格）を設定することができるため、それにより同一の課題に対して、異なるアプローチで課題解決にあたることができる。
・一つの AI 内のキャラクター同士で意見が異なる場合や、対処の難しい事象に遭遇した場合など、AI/AE は時折「酩酊」する。
・AI の「混乱」「酩酊」的表現は、アーサー・C・クラーク『2001 年宇宙の旅』など他の SF でも示されており、高度に進化した AI に起こりえる事象であることが推定される。
・生体的センシングを VR と組み合わせると、人間の仮想的な行動範囲は無限化する。
・視覚、聴覚だけでは、リアルな体験との体感ギャップは依然大きい。（触覚、嗅覚、味覚など、身体的感覚がリアルな体感には重要な要素となる）
・「あざとい」、「だます」は、生物ならではの行動であるので、これには一定程度の知能や知性が必要であることから、人工物の生体感を醸し出す助けの表現であると推定される。
・都市部における「中枢」の場合は、その複雑さ、影響のおよぶ範囲の大きさから、混乱による恐怖を連想させる。
・「AI」の未来社会における役割、使用目的を考えた場合、単なる情報処理にとどまらず、むしろ接触事象に情動し、感情表現をさせることに使うことも有意義となってくる。
・「生体」の一部を人工的なデバイスに置き換えことができれば、人間の機能のマシンへの置き換えは必然の傾向と推定される。
・「AI」=「AE」（人工実在）による対極的意識の共存が予測される。
・生体機能付与による自己進化の可能性があるうる。

4. 考察

現在開発中の先端技術で、前章の解読結果につながる可能性のあるものとして、
・CPU の超微細化によるバイオ・コンピュータ
・モバイル情報の伝達による物理的動作の遠隔化
・生体人工臓器
・「マシン」と「バイオ」の融合
・バイオミメティックス
・有人宇宙飛行
・量子コンピュータ
・運転者の状態、感情を把握する運転支援技術
・ロボット同士のネットワークの成立
などがあげられる。

これらの技術の進歩の先には、先述した解読結果において推測されたような AI の「意識」発

現、もしくは人間とマシンの相互意識の接続はありうるのだろうか。一方、現状の「自動運転車輌」「コミュニケーション・ロボット」「ヒューマノイド・ロボット」などをみれば、今はまだマシンが人間（生体）に近づこうとすればするほど、人間の心身の優位性、複雑性を認識せざるをえない段階にあると言わざるをえない。

この対立した推定に一定の回答を得るには、今回の作品のテキストマイニングをさらに継続し、小松作品の『復活の日』、『果てしなき流れの果に』、『継ぐのは誰か』についても、テキストマイニングによる同様の解析を行う必要があると考えられる。

以上により今回、「未来社会」像の提示ということでは、この結果からはまだ不十分だと言わざるをえない。しかし、この研究方法の方向性は、得られた解読結果と、現実の技術開発の方向およびその社会的影響との比較から、間違ってはいないと言えよう。少なくともこの方法が社会調査の新たな方法としての価値はある。

テキストマイニングによる原著者の発想の外在化と、そこから実際の未来を予測しようとする筆者らの試みが一定の有効性をもちうることについては、今回ある程度確信がもてたが、それを実証するひとつの方法として、この結果をインターネットなどで公開し、広く意見を求めることも必要であろう。

今回はこの方法の一定の有効性が明らかになったので、今後より高速の計算機およびより進歩したテキストマイニングのソフトウエアを使えば、同様の方法でより大量のデータの解析から結論をえることがきる。したがって、本研究は、新たな社会調査とその解読の方法の提示としたい。

付記

本研究は、ヤマハ発動機株式会社の委託による 2017 年度産学連携共同研究による成果の一部である。

参考文献

(1) クライン・ユーベルシュタイン，1980,『SF 思考のすすめ』講談社.
(2) ケヴィン・ケリー（服部桂訳），2014,『TECHNIUM・・・テクノロジーはどこに向かうのか』みすず書房.
(3) 小松左京，1987,『虚無回廊』徳間書房.
(4) 仁平典宏・藤田真文，2017,「特集「テキストマイニングをめぐる方法論とメタ方法論」によせて」『社会学評論』Vol.63, No.3，日本社会学会.

Decoding the “Future Society” by Text Mining

Ioki Hara, Keitaro Okuno,
Takuji Okuno

Abstract

This study confirms the usefulness of the text mining software, which enables the utilization of a massive amount of text data as a new social science tool alternative. It is an alternative for the existing social survey methods used for forecasting the “future society.” It also reports confirmation that the trial use of the method is applicable to a certain degree.

In this study, “Kyomukairō (The Corridor of Emptiness)” - a novel by Sakyo Komatsu - was used for the text mining. The study chose this novel because it is based on solid scientific facts, and also because “AI” and “Man-Machine Interaction”, both frequently featured throughout the novel, are assumed to have a defining impact on our future society. Using keywords related to these two technologies as the core, the study mined the work and extracted keywords indicating the social statuses of these technologies and their involvement in people’s lives in the novel’s imaginary society and attempted to visualize how our “future society” would look like through brainstorming sessions.

While the result of this single study cannot be claimed sufficient in terms of presenting an image of the “future society”, it nevertheless has demonstrated the appropriateness and efficiency of text mining as a study method.

Key words : text mining, future society, social survey method